Assessing the Feasibility of Large-Scale Natural Language Processing in a Corpus of Ordinary Medical Records: A Lexical Analysis

William R. Hersh, M.D.
Emily M. Campbell, R.N., M.S.
Susan E. Malveau, M.S.
Division of Medical Informatics and Outcomes Research
Oregon Health Sciences University

***Objective**: Identify the lexical content of a large corpus of ordinary medical records to assess the feasibility of large-scale natural language processing.*
***Methods**: A corpus of 560 megabytes of medical record text from an academic medical center was broken into individual words and compared with the words in six medical vocabularies, a common word list, and a database of patient names. Unrecognized words were assessed for algorithmic and contextual approaches to identifying more words, while the remainder were analyzed for spelling correctness.*
***Results**: About 60% of the words occurred in the medical vocabularies, common word list, or names database. Of the remainder, one-third were recognizable by other means. Of the remaining unrecognizable words, over three-fourths represented correctly spelled real words and the rest were misspellings.*
***Conclusions**: Large-scale generalized natural language processing methods for the medical record will require expansion of existing vocabularies, spelling error correction, and other algorithmic approaches to map words into those from clinical vocabularies.*

INTRODUCTION

Although a great deal of information about patients is accessible in the electronic medical record (EMR), much information remains "locked" in narrative text [1]. The ability to extract information from this text by computer would be valuable for both clinical care and research, allowing access to data much richer than the diagnosis codes, laboratory data, and fiscal data that is currently used [2]. Unfortunately, large-scale natural language processing (NLP) from ordinary clinical text has been difficult, as a number of problems prevent generalizability and scalability, from language idioms to misspellings [3].

Successful NLP requires two broad capabilities: (a) algorithms to parse text into syntacic and semantic categories and (b) vocabularies to serve as "targets" to allow normalization and codification of the parsed text. In focused domains, researchers have shown success in both categories. A number of investigators have been able to develop approaches that work in specific domains with "cleansed" text [4-10]. Likewise, other investigators have shown the ability to map text into controlled voabularies such as SNOMED [11] and ICD-9 [12]. These studies show that in focused domains, algorithms can be developed to achieve 80-90% recognition of important concepts.

But if NLP is to play a significant role in unlocking information from the EMR, then it must operate on a much larger scale than current implementations. It must also be able to handle the "nuances" of ordinary text, such as document headers, typists' initials, misspellings, and so forth. If these problems are not handled effectively, then the intensive person-hours required to build the types of systems cited above will have difficulty justifying their costs. In this study, we attempted to assess the feasibility of NLP from a large corpus of ordinary on-line clinical narratives by performing a lexical analysis to determine if the words used in the records occurred in existing resources of medical and general terminology. If the words used in text are not part of general lexicons, then reaping the benefits of generalizing the normalization and codification of EMR text will prove difficult.

Another goal of this study was to determine the nature of unrecognizable words, including those not in medical vocabularies or common word lists, which could be either misspellings or medical words not occurring these resources. We also assessed the coverage of words that occurred in other vocabularies but not the UMLS Metathesaurus. This was done because although the UMLS Metathesaurus is not a comprehensive clinical vocabulary, it has been proposed as the foundation of one [13]. If the Metathesaurus will serve in this role, then it is important to know what it does not cover. Previous work showed that its phrasal coverage is incomplete [14].

METHODS

A corpus of 560 megabytes was extracted from the Oregon Health Sciences University (OHSU) EMR. This represented all dictated reports – discharge summaries, radiology reports, progress notes, emergency room reports, and letters – that were entered into the system in 1995. This corpus represented ordinary medical records, with usual procedures for dictation and transcription, according to the director of medical records (personal communication, Jeanne Kistner, OHSU director of medical records). In a previous experiment, OHSU-transcribed medical records were found to have about the same number of words unrecognizable to a medical dictionary as those from four other

1091—8280/97/$5.00 © 1997 AMIA, Inc.

geographically disparate institutions: Brigham and Women's Hospital, LDS Hospital, Columbia-Presbyterian Medical Center, and Stanford Medical Center [15].

Each report in the corpus was tokenized into individual words, where a word was defined as any sequence of alphanumeric characters delimited by white space (most punctuation, including hyphens). All tokens containing embedded colons surrounded by up to three characters on each side (likely to represent typists' initials) and metacharacters were discarded, with the remainder designated as true words, which were then normalized using the UMLS *norm* routine [16].

The normalized words were then compared against the words from all of the terms in six medical vocabularies (which had also been reduced to normalized form):
1. The UMLS Metathesaurus, 1996 [17].
2. The Medical Entities Dictionary, 1995 [18].
3. SNOMED, version 3.2 [19].
4. The Medical Letter Drug List, 1996.
5. Stedman's Medical Dictionary, 27th Edition, 1996.
6. Medical Abbreviations, 1993 [20].

After medical vocabulary words were categorized, the remaining words were compared against two additional word lists:
1. Common words from the Unix spell checker.
2. The database of names from the OHSU EMR system.

At this point, all words from medical vocabularies, a common spell checker, and the database of patient names from which the records were derived had been categorized. The remaining words were analyzed to develop algorithmic processes that could convert them to words from the previously described resources. Those words left after this process were then converted into a keyword-in-content (KWIC) file that listed the two words to the left and the two words to the right for each instance in the corpus. This approach has been used to recognize names in medical records for obfuscation purposes [21]. A series of rules were developed to recognize words in specific patterns.

The algorithmic and contextual recognition left a list of remaining words that were unidentified and represented either spelling errors or words not recognized by any of the above word lists or processes. A 10% sample of these words (every 10th word) was assessed manually by looking at each word's occurrence in the KWIC file and then assigning it of the following categories:
1. A correctly spelled word.
2. A probable correctly spelled word (recognized in context to be a name or product).
3. A misspelling.
4. A "garbage" word, consisting of a string of unrecognized characters.
5. Unable to determine.

After the categorization of words, a final analysis attempted to qualitatively judge the significance of words that:
1. Occurred in other medical vocabularies but not the UMLS Metathesaurus.
2. Did not occur in any vocabulary but had medical significance.

RESULTS

The corpus of text contained 238,898 documents, which yielded a total of 124,993 unique tokens. The average word occurred 613.9 times in 311.6 documents. Table 1 lists the successive mapping of the words into vocabulary categories. The largest category of words was those which occurred in one of the six medical vocabularies. Table 2 lists the proportion of words in each vocabulary that occurred in one or more of the medical vocabularies. The UMLS Metathesaurus had the highest coverage of all vocabularies, with over 80% of words occurring in any medical vocabulary being present in the Metathesaurus.

The next largest proportion of words occurred in the list of common words or patient names. These groups of words were not only the largest, but also those which occurred in the most documents and with the highest frequency. Thus, recognizable medical words, common words, or names comprised the bulk of total words in the corpus.

Nonetheless, 23.1% of the words did not occur in any categories and could not be algorithmically or contextually converted to such words. Of the otherwise unrecognized words that could be classified (i.e., were not unknown), 77.9% were correctly or probably correctly spelled words, representing medical words not in any of the six medical resources, names, products, and brand names.

Table 3 lists the categories of algorithmic processes developed and used to identify words. The process enabled automated conversion of the 10% of the corpus words to those in the medical vocabularies, names list, or Unix spell checker. Nearly half of the words in this category were compound forms, such as gastroduodenal. Few medical vocabularies contain these words that are used commonly in medical dictations.

The categories for the contextual rules are shown in Table 4. About 6% of all words could be classified according to this approach. Over half of these words were names, recognizable by prefixes (e.g., Mr., Dr.) and suffixes (M.D., R.N.). A variety of diseases and anatomical locations not present in the medical resources were detected here.

Table 1 – Amount and proportion of words in categories with average document and overall frequency occurrence.

Category	Amount (percentage)	Avg. docs.	Avg. freq.
Initials and embedded metacharacters	1,344 (1.1%)	157.7	158.1
In one of six medical vocabularies	42,721 (34.2%)	827.5	1658.5
In names list or Unix spell checker	32,100 (25.7%)	75.6	140.6
Algorithmically recognizable	12,592 (10.1%)	15.0	18.2
Recognizable in context	7,311 (5.8%)	9.1	12.2
Otherwise unrecognized	28,925 (23.1%)		
Correctly spelled real word	12,912* (10.3%)	23.7	28.1
Probably correctly spelled	9,101* (7.3%)	5.8	6.6
Incorrectly spelled	6,171* (4.9%)	2.2	2.4
Garbage word	70* (0.1%)	1.4	1.4
Unknown	671* (0.5%)	1.6	1.7
TOTAL	124,993 (100%)	311.6	613.9

* estimated from 10% sample of otherwise unrecognized words

Table 2 – Proportion of words in medical vocabularies by individual vocabulary

Vocabulary	Num. words (percentage)
UMLS Metathesaurus	34,356 (80.4%)
SNOMED	26,722 (62.6%)
Stedman's Medical Dictionary	24,872 (58.2%)
Medical Entities Dictionary	15,499 (36.3%)
Medical Abbreviations	3,319 (7.8%)
The Medical Letter Drug List	1538 (3.6%)

Table 3 – Categories of algorithmically recognizable words and their occurrence.

Category	Example	Number (percentage)
Age	3yr	92 (0.7%)
Cancer staging	t1n1m1	145 (1.2%)
Compound forms	gastrodudodenal	6258 (49.7%)
Dimension	3x5 (cm)	125 (1.0%)
Dosage	q4hours	79 (0.6%)
Doubled/tripled characters	reccomended	1004 (8.0%)
Double first character	aabdominal	178 (1.4%)
Gases	vO2max	4 (<0.1%)
Gerund forms	conferencing	101 (0.8%)
Gestational age	31w2d	343 (2.7%)
Gravida/para information	G3P2	5 (<0.1%)
IV fluid	D5NS	5 (<0.1%)
Known words with appended numbers	abbreviated12	506 (4.0%)
Length	25mm	56 (0.4%)
Liquid measure	15cc	59 (0.5%)
Noun forms	streakiness	141 (1.1%)
Numbers	63rd	289 (2.3%)
Plural or possessive of name	Emily's	573 (4.6%)
Radiology terms	3view	2 (<0.1%)
Rate	4bpm	6 (<0.1%)
Single characters with numbers	a0983	967 (7.8%)
Specialty clinic names with trailing initials of providers	CardiologyJGY	1306 (10.4%)
Temperature	101F	55 (0.4%)
Time	4wks	96 (0.8%)
Typo - letter L used for number 1	l2th	8 (0.1%)
Typo - number 0 used for letter O	m0nths	44 (0.3%)
Typo - number 1 used for letter L	a1cs	8 (0.1%)
Weight	47lbs	128 (1.0%)

Table 4 – Categories of contextually recognizable words and their occurrence. The underlined term in the example represents the anchor for the context.

Category	Example	Number (percentage)
Names	Dr. William Hersh	4141 (56.6%)
Cities	Beaverton, Oregon	225 (3.1%)
Initials	Cardiology JGY	592 (8.1%)
Streets	Capitol Highway	210 (2.9%)
Places	Lloyd Center	388 (5.3%)
Diseases	Allagiles Syndrome	435 (5.9%)
Medicines	Alimentum mg	544 (7.4%)
Joint descriptions	Calcaneotalar ligaments	86 (1.2%)
Medical tests	Digitract monitoring	118 (1.6%)
Equipment	Accufix suture	227 (3.1%)
Characters/Numbers	22q11 deletion	345 (4.7%)

Table 5 – Medical words in other vocabularies but not in UMLS Metathesaurus.

abruption
bacteriocidal
centimeter
depakote
extranuclear
hemiarthroplasty
interosseus
lozenge
lucid
sonogram

Table 6 – Medical words not in any vocabulary.

ascended
cavernosometry
dipsticked
globulinemia
heplocked
laryngotracheoplasties
malodor
nephroblastomatosis
oculomyasthenia
righthandedness

As already noted, the largest proportion of words originated from the medical vocabularies. Further analysis identified 7,479 words that occurred in one of the other five medical vocabularies but not the UMLS Metathesaurus. A sample of medical words not present in the Metathesaurus is shown in Table 5.

The analysis also showed that approximately 10% of the words represented correctly spelled words not in any vocabulary. Table 6 lists a sample of these words that are medically oriented.

DISCUSSION

About 40% of the words in a large corpus of ordinary medical records do not occur in medical vocabularies, a common word list, or a database of names from which the records were derived. About one-third of these words can be recognized with algorithmic modification and/or contextual rules. Of the remaining non-recognizable words, three-fourths represent correctly spelled real words while the remainder are misspellings. This indicates that large-scale generalized NLP methods for the EMR will require expansion of existing vocabularies, spelling error correction, and other algorithmic approaches to map words into those from clinical vocabularies.

There were some limitations of this study. First, these dictated records come from only one institution. Although there is evidence that dictated reports from other institutions are similar, this is not known for this corpus of records. Second, this analysis assumes that these words have only one sense. That is, a medical word might also be a name, or an abbreviation might be a misspelling. We attempted to minimize this type of error by categorizing medical words first. Finally, while the vocabularies used in this study represent a broad cross-section of available resources, they are not the only ones available. Furthermore, these represent current resources that will no doubt improve over time.

Future plans for this work include expansion of the analysis to records from other sites and with updated versions of the vocabularies. We also plan to share our findings with vocabulary developers so that they may use them to enhance their vocabularies.

ACKNOWLEDGEMENTS

This study was supported by NLM Cooperative Agreement LM05879.

REFERENCES

1. Hripcsak, G., *et al.*, *Unlocking clinical data from narrative reports: a study of natural language processing.* Annals of Internal Medicine, 1995. **122**: p. 681-688.

2. Dolin, R., *Outcome analysis: considerations for an electronic health record.* M.D. Computing, 1997. **14**: p. 50-56.

3. Macleod, C., S. Chen, and J. Clifford, *Parsing Unedited Medical Narrative*, in *Medical Language Processing: Computer Management of Narrative Data*, N. Sager, C. Friedman, and M. Lyman, Editors. 1987, Addison-Wesley: Reading, MA. p. 163-173.

4. Friedman, C., *et al.*, *A general natural-language text processor for clinical radiology.* Journal of the American Medical Informatics Association, 1994. **1**: p. 161-174.

5. Lenert, L. and M. Tovar. *Automated linkage of free-text descriptions of patients with a practice guideline.* in *Proceedings of the 17th Annual Symposium on Computer Applications in Medical Care.* 1993. Washington, DC: McGraw-Hill.

6. Haug, P., *et al. A natural language understanding system combining syntactic and semantic techniques.* in *Proceedings of the 18th Annual Symposium on Computer Applications in Medical Care.* 1994. Washington, DC: Hanley Belfus.

7. Sager, N., *et al.*, *Natural language processing and the representation of clinical data.* Journal of the American Medical Informatics Association, 1994. **1**: p. 142-160.

8. Ertle, A., E. Campbell, and W. Hersh. *Automated application of clinical practice guidelines for asthma management.* in *Proceedings of the 19th Annual AMIA Fall Symposium.* 1996. Washington, DC: Hanley-Belfus.

9. Evans, D., *et al. Automating concept identification in the electronic medical record: an experiment in extracting dosage information.* in *Proceedings of the 19th Annual AMIA Fall Symposium.* 1996. Washington, DC: Hanley-Belfus.

10. Jain, N., *et al. Identification of suspected tuberculosis patients based on natural language processing of chest radiograph reports.* in *Proceedings of the 19th Annual AMIA Fall Symposium.* 1996. Washington, DC: Hanley-Belfus.

11. Sager, N., *et al. Automatic encoding into SNOMED III: a preliminary investigation.* in *Proceedings of the 18th Annual Symposium on Computer Applications in Medical Care.* 1994. Washington, DC: Hanley Belfus.

12. Larkey, L. and W. Croft. *Combining classifiers in text categorization.* in *Proceedings of the 19th Annual International ACM Special Interest Group in Information Retrieval.* 1996. Zurich: ACM Press.

13. Humphreys, B., *et al.*, *Planned NLM/AHCPR large-scale vocabulary test: using UMLS technology to determine the extent to which controlled vocabularies cover terminology needed for health care and public health.* Journal of the American Medical Informatics Association, 1996. **3**: p. 281-287.

14. Hersh, W., *et al. Empirical, automated vocabulary discovery using large text corpora and advanced natural language processing tools.* in *Proceedings of the 19th Annual AMIA Fall Symposium.* 1996. Washington, DC: Hanley-Belfus.

15. Evans, D. and W. Hersh, *The CXR Reports: Model and Analysis*, . 1993, Laboratory for Computational Linguistics, Carnegie Mellon University.

16. McCray, A., *et al.*, *UMLS knowledge for biomedical language processing.* Bulletin of the Medical Library Association, 1993. **81**: p. 184-194.

17. Lindberg, D., B. Humphreys, and A. McCray, *The unified medical language system project.* Methods of Information in Medicine, 1993. **32**: p. 281-291.

18. Cimino, J., *et al.*, *Knowledge-based approaches to the maintenance of a large controlled medical terminology.* Journal of the American Medical Informatics Association, 1994. **1**: p. 35-50.

19. Rothwell, D., *et al. Developing a standard data structure for medical lanaguage - the SNOMED proposal.* in *Proceedings of the 17th Annual Symposium on Computer Applications in Medical Care.* 1993. Washington, DC: McGraw-Hill.

20. Davis, N., *Medical Abbreviations: 8600 Conveniences at the Expense of Communications and Safety.* 1993, Huntingdon Valley, PA: Neil M. Davis Associates.

21. Sweeney, L. *Replacing personally-identifying information in medical records, the Scrub system.* in *Proceedings of the 19th Annual AMIA Fall Symposium.* 1996. Washington, DC: Hanley-Belfus.